Analysis of Frequencies

INTD8065 Data Analysis for Cancer Research
María-Eglée Pérez and Luis Raúl Pericchi

Contents of this class

Analyzing frequencies

In the following, we will focus on the analysis of one or more categorical variables, particularly when we have counts of observations in each combination of the variables (*contingency tables*)
A fundamental statistic for the analysis of categorical data is the (*Pearson*) *chi-square* (χ^2) *statistic*, which is commonly used to compar observed and expected frequencies in categories

$$X^2 = \sum \frac{(o-e)^2}{e}$$

This statistic is compared against a χ^2 reference, whose degrees of freedom depend on the specific problem and is a function of the number of categories minus one.

Null hypotheses in categorical analysis often imply that a sample of observations came from a population where the observed frequencies match some expected frequencies.
The X^2 statistic approximately follows a χ^2 distribution if the following assumptions hold

- Observations are classified into categories independently.
- No more than 20% of the categories have expected frequencies less than about five.

We want to test if our observations come from a population with a particular distribution of frequencies in categories of a single variable. The general data layout for these tests is usually a single categorical variable with counts of frequencies for each category.

Example: Ninety shrubs of a dioecious plant were sampled in a forest and each plant was classified as male or female. 40 females and 50 males were observed. Are these data consistent with the hypothesis of equal proportions of male and females?

For this example, our null hypothesis is $H_0 : p_F = p_M = \frac{1}{2}$. So, our situation is:

	Female	Male	Total
Observed	40	50	90
Expected	45	45	
$o - e$	5	5	
$\frac{(o-e)^2}{e}$	0.556	0.556	

So, $X^2 = 1.112$, and this statistic should be compared with a χ^2 with 1 degree of freedom.

For performing this test in *R*, we can use the command `chisq.test`.

```
> chisq.test(c(40,50),p=c(0.5,0.5))

        Chi-squared test for given probabilities

data:  c(40, 50) X-squared = 1.1111, df = 1, p-value = 0.2918
```

The null hypothesis can't be rejected

Two Way Contingency Tables

The general form of a two way contingency table is

	Category B				
Category A	B_1	B_2	$\ldots$	B_b	Total
A_1	n_{11}	n_{12}	$\ldots$	n_{1b}	$n_{1.}$
A_2	n_{21}	n_{22}	$\ldots$	n_{2b}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_a	n_{a1}	n_{a2}	$\ldots$	n_{ab}	$n_{a.}$
Total	$n_{.1}$	$n_{.2}$	$\ldots$	$n_{.b}$	$n_{..}$

The null hypothesis in this situation is H_0 : The factors of row and column are independent. The χ^2 statistic corresponding to this test is

$$X^2 = \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

where

$$E_{ij} = \frac{n_{i.} n_{.j}}{n_{..}}$$

The reference distribution for this test is a χ^2 with $(a-1)(b-1)$ degrees of freedom.

Example:

The following data correspond to the eye color and hair color of 5387 children in Caithness, Scotland. We want to determine if there exists association between these two variables.

	Hair				
Eye	Yellow	Red	Medium	Dark	Black
Blue	326	38	241	110	3
Light	688	116	584	188	4
Medium	343	84	909	412	26
Dark	98	48	403	681	85

```
> haireye.tab=matrix(scan("haireye.dat"),ncol=5,byrow=T)
Read 20 items
> haireye.tab
     [,1] [,2] [,3] [,4] [,5]
[1,]  326   38  241  110    3
[2,]  688  116  584  188    4
[3,]  343   84  909  412   26
[4,]   98   48  403  681   85
> chisq.test(haireye.tab)

        Pearson's Chi-squared test

data:  haireye.tab
X-squared = 1240.039, df = 12, p-value < 2.2e-16
```

The null hypothesis is rejected, and we conclude that there is a relationship between both factors.

Generalized Linear Models

The Generalized Linear Models theory extends partially the results for the Normal Linear Model to situations where the involved distributions are not normal, but share some of its characteristics (*Exponential Family of distributions*).

Particular cases of Generalized Linear Models include methods as logistic models, log-linear models, some cases of survival analysis, etc. (and of course the normal linear model).

Least square estimation no longer applies and maximum likelihood methods must be used. Also, reference distributions for hypothesis testing are not exact, but rather approximations to the real distributions.

A GLM consists of three components

1. Response variables $Y_1, \ldots, Y_n$ which follow distributions in the exponential family of distributions, which includes normal, binomial, Poisson, gamma and negative binomial. Probability distributions from the exponential family of distributions can be defined by the natural parameter, a function of the mean. Each Y_i depends of the natural parameter θ_i (the θ_i can be different).

$$f(y_i; \theta_i) = \exp\{y_i b_i(\theta_i) + c_i(\theta_i) + d_i(y_i)\}$$

2. A set of parameters

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

and explanatory variables

$$X = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix}$$

3. A monotonic function, called *link function*, such that

$$\eta = g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$$

where

$$\mu_i = E(Y_i)$$

Common link functions include

1. Identity link, which is $g(\mu) = \mu$, and models the mean or expected value of Y. This is used in normal linear models.
2. Log link, which is $g(\mu) = \log(\mu)$ and models the log of the mean. This is used for count data (which cannot be negative) in log-linear models.
3. Logit link, which is $g(\mu) = \log \frac{\mu}{1-\mu}$, and is used for binary data and logistic regression.

As we said before, the parameters of the model are estimated using maximum likelihood.
Approximated tests for the null hypothesis $H_0 : \beta_i = 0$ vs $H_i : \beta_i \neq 0$ are based on the fact that, under the null hypothesis (for big samples),

$$z_i = \frac{\hat{\beta}_i}{\sqrt{v_{ii}}} \sim N(0,1)$$

where v_{ii} is the i-th term in the diagonal of matrix I^{-1}, where I is the Fisher information matrix

$$I_{ij} = E\left(-\frac{\partial^2 \log L}{\partial\beta_i \partial\beta_j}\right)\Bigg|_{\boldsymbol{\beta}=\mathbf{b}}.$$

We reject H_0 when $|z_i| > z_{\alpha/2}$.

Tests of the goodness of fit of a model are based on the *Deviance*, which is defined as

$$D = 2 \log \lambda = 2[l(\mathbf{b}_{\max}; \mathbf{y}) - l(\mathbf{b}; \mathbf{y})]$$

Here λ is the likelihood ratio statistic for comparing the *maximal model* (a model of the same family with the maximum number of parameters) and the model of interest. A small deviance corresponds to a model which is fitting well the data.
If the model is "good", it can be shown that, approximately,

$$D \sim \chi^2_{n-p}$$

So, we'll reject the model in favor of the maximal model if D is big enough (that is, if $D > \chi^2_{n-p}(\alpha)$)

The deviance can also be used for comparing GLM's. Suppose that $M_0 \subset M$ are nested models with q and p parameters respectively ($q < p$). Let D_{M_0} and D_M be their deviances. Then, if the null hypothesis that the simplest model is correct is true,

$$D_{M_0} - D_M \sim \chi^2_{p-q}$$

This distribution is approximate, and it is exact only for normal errors models.

The definition of AIC (Akaike information criterium) can be extended naturally to GLM's

$$AIC = -2 \text{ max of the log-likelihood} + 2p$$

where p is the number of parameters for the model.
Generalized linear models can be fitted in *R* using the command `glm`. Models can be again compared using the command `anova`. We'll see in some examples how to use these commands.

Generalized Linear Models for Binary Variables

We will consider the following type of response variable

$$Z = \begin{cases} 1 & \text{if we get a success} \\ 0 & \text{if we get a failure} \end{cases}$$

with $P(Z = 1) = \pi$ and $P(Z = 0) = 1 - \pi$. If we have n of such variables $Z_1, \ldots, Z_n$, and they are independent with, $P(Z_i = 1) = \pi_i$, then their joint probability is in the exponential family

$$\prod_{j=1}^{n} \pi^{z_j}(1-\pi)^{1-z_j} = \exp\left[\sum_{j=1}^{n} z_j \log\left(\frac{\pi_j}{1-\pi_j}\right) + \sum_{j=1}^{n} \log(1-\pi_j)\right]$$

If all the π_j are equal, let Y be the number of successes in n trials.

$$Y = \sum_{j=1}^{n} Z_j$$

The distribution of Y is Binomial(n, π). This distribution is also in the exponential family.

Finally, lets consider the case of N independent variables $Y_1, Y_2, \ldots, Y_N$, which correspond to the number of successes in N subgroups o strata.

	Subgroups			
	1	2	...	N
Successes	Y_1	Y_2	...	Y_N
Failures	$n_1 - Y_1$	$n_2 - Y_2$	...	$n_N - Y_N$
Totals	n_1	n_2	...	n_N

If $Y_i \sim \text{Bin}(n_i, \pi_i)$, the log-likelihood is

$$l(\pi_1, \ldots, \pi_N; y_1, \ldots, y_N)$$

$$= \sum_{j=1}^{N} \left[y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + n_i \log(1-\pi_i) + \log\binom{n_i}{y_i} \right]$$

For any of these three situations, a generalized linear model can be fitted, given a convenient link function.

Link Functions

We wish to describe the proportion of successes $P_i = y_i/n_i$ in each subgroup in terms of the levels of a factor or of explanatory variables characterizing the subgroup. This will be done modelling the probabilities pi_j as

$$g(\pi_j) = \mathbf{x}_i^T \boldsymbol{\beta}$$

It is usual to employ as link function the inverse of a probability distribution to guarantee that π is in $[0, 1]$

$$\pi = g^{-1}(\mathbf{x}^T \boldsymbol{\beta}) = \int_{-\infty}^{t} f(s)ds$$

where $f(s) \geq 0$ and $\int_{-\infty}^{\infty} f(s)ds = 1$.

Some link functions:

- **Probit link**
  The link function is the inverse of the normal(0,1) distribution.

$$\pi = \Phi(\mathbf{x}^T \boldsymbol{\beta})$$

  Equivalently, $g = \Phi^{-1}$, and so

$$\Phi^{-1}(\pi) = \mathbf{x}^T \boldsymbol{\beta}$$

  Probit models are used in some areas in biological sciences and social sciences, where they have a natural interpretation.

- **Complementary log log link**

$$\pi = 1 - \exp[-\exp(\mathbf{x}^T \boldsymbol{\beta})]$$

  Equivalently,

$$\log[-\log(1 - \pi)] = \mathbf{x}^T \boldsymbol{\beta}$$

- **Logistic link**
  This is the most used link, and the one we will concentrate on.

$$\pi = \frac{\exp(\mathbf{x}^T\boldsymbol{\beta})}{1+\exp(\mathbf{x}^T\boldsymbol{\beta})}$$

  This is equivalent to

$$\log\left(\frac{\pi}{1-\pi}\right) = \mathbf{x}^T\boldsymbol{\beta}$$

Odds and Odds Ratio

Consider a 2×2 table

	Subgroups	
	1	2
Successes	Y_1	Y_2
Failures	$n_1 - Y_1$	$n_2 - Y_2$
Totals	n_1	n_2

Let π_1 y π_2 be the probabilities of success for groups 1 and 2, respectively. We will define the *odds* for group i as

$$O_i = \frac{\pi_i}{1 - \pi_i}$$

i.e., the odds are the rate between the probability of success and the probability of failure. A logistic model assigns a linear structure to the logarithm of the odds for each group.

The *"odds ratio"* between both categories is defined as

$$OR = \frac{\frac{\pi_2}{1-\pi_2}}{\frac{\pi_1}{1-\pi_1}} = \frac{\pi_2(1-\pi_1)}{\pi_1(1-\pi_2)}$$

The odds ratio is used as a measure of association between rows and columns in the table. If it is close to 1, both groups have the same distribucion of successes and failures. If $OR > 1$, group 2 has a higher probability of success, and viceversa if $OR < 1$.

Consider a simple model for this situation

$$\log\left(\frac{\pi_j}{1-\pi_j}\right) = \beta_0 + \beta_1 x_j$$

Suppose first that x is an indicator variable, with value 0 for group 1 and 1 for group 2. Then

$$\log(O_1) = \beta_0$$

$$\log(O_2) = \beta_0 + \beta_1$$

So

$$OR = \exp(\log(O_2) - \log(O_1)) = e^{\beta_1}$$

That is, β_1 is the logarithm of the Odds Ratio, and so it is a measure of association between rows and comlumns.

If x is a numerical variable, then

$$\log(O_x) = \beta_0 + \beta_1 x$$

$$\log(O_{x+1}) = \beta_0 + \beta_1 (x+1)$$

and the odds ratio is

$$\exp(\log(O_{x+1}) - \log(O_x)) = e^{\beta_1}$$

In this case, β_1 represents the odds ratio associated with the increment of the explanatory variable in one unit.

Consider now the model

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i}$$

Suppose that both x_1 and x_2 are indicator variables (absence or presence). Then

1. If x_2 is present

$$O_{01} = e^{\beta_2}$$

$$O_{11} = e^{\beta_1+\beta_2+\beta_3}$$

And so the quotient of this quantities $e^{\beta_1+\beta_3}$, is the odds ratio for x_1 in presence of x_2.

2. If x_2 is absent,

$$O_{00} = 1$$

$$O_{10} = e^{\beta_1}$$

and e^{β_1}, represents the odds ratio for x_1 when x_2 is absent.

This means that the effect of x_1 depends on the level of x_2 (and viceversa). This situation is called *effect modification*, and is represented by the introduction of interaction terms in the model.

Example

When a patient is diagnosed as having cancer of the prostate, an important question in deciding on treatment strategy for the patient is whether the cancer has spread to the neighboring lymphnodes.. The question is so critical in prognosis and treatment that it is customary to operate on the patient (i.e. perform a laparotomy) for the sole purpose of examining the nodes and removing tissue samples to examine under the microscope for evidence of cancer. However, certain variables that can be measured without surgery are predictive of the nodal involvement, Data in file `prostate.dat` correspond for 53 prostate cancer patients receiving surgery, and we sant to determine which of five preoperative variables are predictive of nodal involvement.

These variables are

1. X ray reading (`Xray`)
2. Result of a pathological analysis of a biopsy (`Grade`)
3. Stage of the tumour obtained by palpation with the finger via the rectum. (`Stage`: 1= Positive finding, 0= Negative finding).
4. Age at diagnosis (`Age`)
5. Level of serum acid phosphatase ($\times$100, called `Acid`)

The response variable is the finding at surgery (1=nodal involvement, 0=no nodal involvement)

```
> prostate.frm=read.table("prostate.txt",header=T)
> attach(prostate.frm)
> prostate.mod1=glm(Nodes~Xray+Grade+Stage+Age+Acid,family=binomial)
> summary(prostate.mod1)

Call:
glm(formula = Nodes ~ Xray + Grade + Stage + Age + Acid,
family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0110  -0.7021  -0.3654   0.5723   1.9852

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.06180    3.45992   0.018   0.9857
Xray         2.04534    0.80718   2.534   0.0113 *
Grade        0.76142    0.77077   0.988   0.3232
Stage        1.56410    0.77401   2.021   0.0433 *
```

```
Age          -0.06926    0.05788  -1.197   0.2314
Acid          0.02434    0.01316   1.850   0.0643 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 70.252  on 52  degrees of freedom
Residual deviance: 48.126  on 47  degrees of freedom
AIC: 60.126

Number of Fisher Scoring iterations: 5

> 1-pchisq(48.126,47)
[1] 0.4270388
```

The model performs equivalently to the saturated model, and so it doesn't seem necessary to include interactions between the variables.
Lets try to simplify the model

```
> drop1(prostate.mod1,test="Chisq")
Single term deletions

Model:
Nodes ~ Xray + Grade + Stage + Age + Acid
       Df Deviance    AIC    LRT  Pr(Chi)
<none>      48.126 60.126
Xray    1   55.350 65.350  7.224 0.007195 **
Grade   1   49.097 59.097  0.972 0.324263
Stage   1   52.558 62.558  4.432 0.035263 *
Age     1   49.615 59.615  1.490 0.222267
Acid    1   51.572 61.572  3.446 0.063413 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The variable to be eliminated seems to be Grade

```
> prostate.mod2=glm(Nodes~Xray+Stage+Age+Acid,family=binomial)
> summary(prostate.mod2)

Call:
glm(formula = Nodes ~ Xray + Stage + Age + Acid, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8713  -0.6968  -0.3935   0.6053   1.9870

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.44600    3.41443   0.131  0.89607
Xray         2.09770    0.79510   2.638  0.00833 **
Stage        1.76400    0.74686   2.362  0.01818 *
Age         -0.07025    0.05742  -1.224  0.22113
```

```
Acid          0.02218    0.01278   1.735  0.08277 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 70.252  on 52  degrees of freedom
Residual deviance: 49.097  on 48  degrees of freedom
AIC: 59.097

Number of Fisher Scoring iterations: 4

> 1-pchisq(49.097,48)
[1] 0.4289294
```

Again, this simpler model fits well the data, and we'll try to simplify it even more.

```
> drop1(prostate.mod2,test="Chisq")
Single term deletions

Model:
Nodes ~ Xray + Stage + Age + Acid
       Df Deviance    AIC    LRT  Pr(Chi)
<none>      49.097 59.097
Xray    1   57.016 65.016  7.918 0.004894 **
Stage   1   55.381 63.381  6.284 0.012183 *
Age     1   50.660 58.660  1.562 0.211347
Acid    1   52.085 60.085  2.988 0.083894 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The variable that will be removed is Age.

```
> prostate.mod3=glm(Nodes~Xray+Stage+Acid,family=binomial)
> summary(prostate.mod3)

Call:
glm(formula = Nodes ~ Xray + Stage + Acid, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8630  -0.8508  -0.3889   0.5721   2.2386

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.57565    1.18115  -3.027  0.00247 **
Xray         2.06179    0.77767   2.651  0.00802 **
Stage        1.75556    0.73902   2.376  0.01752 *
Acid         0.02063    0.01265   1.631  0.10291
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 70.252  on 52  degrees of freedom
Residual deviance: 50.660  on 49  degrees of freedom
AIC: 58.66

Number of Fisher Scoring iterations: 4

> 1-pchisq(50.660,49)
[1] 0.4078615
```

This simpler model still fits well the data.

```
> drop1(prostate.mod3,test="Chisq")
Single term deletions

Model:
Nodes ~ Xray + Stage + Acid
       Df Deviance    AIC    LRT  Pr(Chi)
<none>      50.660 58.660
Xray    1   58.613 64.613  7.954 0.004798 **
Stage   1   57.059 63.059  6.399 0.011418 *
Acid    1   53.353 59.353  2.694 0.100740
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

It is still possible to remove the variable `Acid`

```
> prostate.mod4=glm(Nodes~Xray+Stage,family=binomial)
> summary(prostate.mod4)

Call:
glm(formula = Nodes ~ Xray + Stage, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9166  -0.9907  -0.4934   0.5892   2.0815

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.0446     0.6100  -3.352 0.000802 ***
Xray          2.1194     0.7468   2.838 0.004541 **
Stage         1.5883     0.7000   2.269 0.023274 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 70.252  on 52  degrees of freedom
Residual deviance: 53.353  on 50  degrees of freedom
AIC: 59.353

Number of Fisher Scoring iterations: 4

> 1-pchisq(53.353,50)
[1] 0.3466239
```

As a final check, we can compare our last model with the first model we fitted

```
> anova(prostate.mod4,prostate.mod1,test="Chisq")
Analysis of Deviance Table

Model 1: Nodes ~ Xray + Stage
Model 2: Nodes ~ Xray + Grade + Stage + Age + Acid
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        50     53.353
2        47     48.126  3    5.228     0.156
```

With base in the model, we can predict the probabilities of the involvement of the lymphatic nodes for all possible combinations of level of `Xray` and `Stage`.

```
> predict(prostate.mod4,data.frame(Xray=c(0,0,1,1),
+ Stage=c(0,1,0,1)),type="response")
        1         2         3         4
0.1145964 0.3878515 0.5186952 0.8406524
```

The highest probability of node involvement corresponds to those patients with positive X rays reading and positive result in the pathological examination of the biopsy.

Example

The following data come from a study of a vaccine against rotavirus diarrheas in children. The table shows the number of cases of diarrhea in children from different socioeconomic groups (measured through the Graffar score) for the vaccinated and not vaccinated groups. We want to know if there is association between the effect of the vaccine and the socioeconomic status.

Graffar	Total vaccine	R+ vaccine	Total placebo	R+ placebo
2-3	382	22	407	39
4-5	721	48	675	94

Here the presence of rotavirus positive diarrheas can be considered as the response variable, and we want to know if the vaccine and the socioeconomic level affect the probability of suffering a rotavirus positive diarrhea.

Note that this example differs of the previous one because the response comes in terms of number of successes, instead of 0 - 1.
In this case, the response variable is expressed by means of a matrix whose first column is the number of successes and whose second column is the number of failures.

```
> rota.pos=c(22,48,39,94)
> totals=c(382,721,407,675)
> rota.neg=totals-rota.pos
> vacuna.fac=gl(2,2,labels=c("vaccine","placebo"))
> graffar.fac = gl(2,1,4,labels=c("2-3","4-5"))
```

For testing the hypothesis of effect modification, we will fit the additive model and we will compare it against the saturated model.

```
> vaccine.mod1=glm(cbind(rota.pos,rota.neg)~vaccine.fac+graffar.fac,
+ family=binomial)
> summary(vaccine.mod1)

Call: glm(formula = cbind(rota.pos, rota.neg) ~ vaccine.fac +
graffar.fac,
    family = binomial)

Deviance Residuals:
      1         2         3         4
 0.5418   -0.3504   -0.3939    0.2666

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -2.9162     0.1689 -17.262  < 2e-16 ***
vaccine.facplacebo  0.7374     0.1546   4.770 1.84e-06 ***
graffar.fac4-5      0.3276     0.1609   2.036   0.0418 *
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 28.16316  on 3  degrees of freedom
Residual deviance:  0.64245  on 1  degrees of freedom
AIC: 28.801

Number of Fisher Scoring iterations: 3

> 1-pchisq(0.64245,1)
[1] 0.422825
```

According to this test, the fitted model is adequate for explaining the data, and we can't reject the hypothesis of absence of interaction; so, the effect of the vaccine is not altered by socioeconomic level. Additionally, note that the normal tests are significant for both factors, and so we can say that the probability of suffering a rotavirus positive diarrhea episode is affected both by the vaccine and the socioeconomic level.

Interpretation of the coefficients:

1. The odds ratio corresponding to the vaccine is $e^{0.7374} = 2.09$. This means that a baby of the placebo groups has a higher probability of suffering a rotavirus positive diarrhea (approximately twice the *risk* of suffering a rotavirus positive diarrhea)
2. The odds ratio corresponding to graffar level is $e^{0.3276} = 1.39$, and we conclude that a child with graffar level 4-5 has 1.39 times a higher risk of suffering a rotavirus positive diarrheal episode.

We can predict the probability of suffering a rotavirus positive diarrhea for each group

```
> cbind(as.character(vaccine.fac),as.character(graffar.fac),
+ round(predict(vaccine.mod1,type="response"),4))
  [,1]      [,2]  [,3]
1 "vaccine" "2-3" "0.0514"
2 "vaccine" "4-5" "0.0699"
3 "placebo" "2-3" "0.1017"
4 "placebo" "4-5" "0.1357"
```

These predicted values confirm the previous affirmations.